DOCUMENT RESUME

ED 427 698 IR 019 249

AUTHOR	Doan, Bich-Lien; Beigbeder, Michel; Girardot, Jean-Jacques; Jaillon, Philippe
TITLE	Using Metadata To Improve Organization and Information Retrieval on the WWW.
PUB DATE	1998-11-00
NOTE	9p.; In: WebNet 98 World Conference of the WWW, Internet, and Intranet Proceedings (3rd, Orlando, FL, November 7-12, 1998); see IR 019 231. Figures may not reproduce clearly.
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Access to Information; Artificial Intelligence; Computer System Design; Foreign Countries; *Indexing; *Information Retrieval; Relevance (Information Retrieval); *Robotics; Search Strategies; User Needs (Information); *World Wide Web
IDENTIFIERS	Dublin Core; France; Links (Indexing); *Metadata; Search Engines; Web Pages; Web Sites

ABSTRACT

The growing volume of heterogeneous and distributed information on the World Wide Web has made it increasingly difficult for existing tools to retrieve relevant information. To improve the performance of these tools, this paper suggests how to handle two aspects of the problem. The first aspect concerns a better representation and description of Web pages. A new concept of Web documents described by metadata is introduced; Dublin Core semantics and the XML syntax are used to represent these metadata. Ways that this concept can improve information retrieval on the Web and reduce the network load generated by robots are suggested. The second aspect is a flexible architecture based on two kinds of robots--"generalists" and "specialists" that collect and organize metadata, in order to localize the resources on the Web. The robots will contribute to the overall auto-organizing information process by exchanging their indices. Two figures present possible organizations of Web pages and interaction between specialists and generalists to answer a user need. (Author/AEF)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Using Metadata to Improve Organization and Information Retrieval on the WWW

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
G. H. Marks
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Bich-Liên Doan, Michel Beigbeder, Jean-Jacques Girardot, Philippe Jaillon
Dpt. RIM, École des Mines de Saint-Etienne
158, cours Fauriel, 42023 Saint-Etienne, France
e-mail: {doan, mbeig, girardot, jaillon}@emse.fr

ED 427 698

Abstract: Until now the growing volume of heterogeneous and distributed information on the WWW makes increasingly difficult for the existing tools to retrieve relevant information.
To improve the performance of these tools, we suggest to handle two aspects of the problem:
One concerns a better representation and description of WWW pages, we introduce here a new concept of "WWW documents", and we describe them thanks to metadata. We'll use the Dublin Core semantics and the XML syntax to represent these metadata. We'll suggest how this concept can improve information retrieval on the WWW and reduce the network load generated by robots.
Then, we describe a flexible architecture based on two kinds of robots : "generalists" and "specialists" that collect and organize these metadata, in order to localize the resources on the WWW. They will contribute to the overall auto-organizing information process by exchanging their indices.

1. Introduction

On the WWW, many search tools are now available to help users access information more easily. But these tools are giving often irrelevant responses as they do not focus on the particular context expected by the user.
What do we mean by context?
Suppose that a user wants to know ten very well-known sites dealing with *environment*. He may ask with:
`Q1 = environment`
The answer to query Q1 contains many correct matches, but they are lost in a lot of noise. AltaVista gives back some 3 000 000 responses matching the word environment, containing pages dealing with computers, ecology and other topics. These different answers are embedded within different conceptual domains, which are part of their respective contexts. But the contexts are generally not explicit in the inqueries, nor they are recognized during the flat indexing done by usual robots. Moreover, even if the result contains pages dealing with environment, these pages are not organized, i.e. we cannot have a resumé of the sites which contains relevant information, inducing difficulty to exploit the responses.
Suppose that another user looks for information on *water treatment* . He can try the query:
`Q2 = water treatment`
Q2 induces noise, because *water treatment* occurs within both the medicine and the environment domains.
So, he can try to refine his query by adding the contextual word *environment*, leading to query Q3:
`Q3 = water treatment + environment`
This time, he or she would get silence because pages containing only *water treatment* and not *environment* are not retrieved. In this case, the context *environment* was implicit and did not appear on the pages, but was explicit in the query.

IR019249

BEST COPY AVAILABLE

To help capture the semantics associated with each site, we propose to redefine the concept of a WWW document as an abstraction of a set of page which can be organized into different hierarchies of clusters. We introduce metadata that describes the documents at an apt level of granularity.

Then we propose to define a flexible architecture using the existing and available search tools on the WWW which will enable everybody to participate in the improvement of document descriptions. This architecture is based upon high interactivity between search tools and progressive organization of information on the WWW. With minimal effort we can use cooperation between existing isolated elements of the WWW, resolving the problems of scalability of centralized servers, networked bandwidth overload, while improving the quality of information retrieval on the WWW by reducing noise and silence.

2. Related work

At present, information retrieval on the WWW is made by search tools that have limited capabilities. Two kinds of tools are used currently :

2.1 Universal robots

These robots are "universal" in the sense that they try to index the whole Web, no matter what topics are addressed by the pages, or where the pages are localized. As the WWW grows, universal robots become more numerous, resulting in overload network bandwidths. They become inadequate in finding relevant information all over the WWW. Their main weaknesses are:

* too many irrelevant responses,
* no organization in the responses leads to difficult way to exploit them.
* loss of context around the responses because of the lack of semantic description while indexing and the lack of expressive request (terms with boolean operators),
* no access to the non-textual documents (images, sounds, video) which are not indexed.

2.2 Thematic directory

Other search tools exist, such as Yahoo, which provide users with the means to browse a hierarchy of thematic directories. A provider can register by filling a form to describe its site, indicating in which topic he wishes his server to appear and at which particular level in the tree of subjects. The advantages of this process is to enable exploratory research and better control in indexing (reducing the noise). The problems encountered by this approach are:

* manual indexing,
* only a part of WWW is indexed, so there is lot of silence in the answers,
* manual classification of the universal information requires manual and high cost for maintenance and updating,
* no content-text indexing of pages.

2.3 Metadata

The use of the tag META in HTML page has been solicited. The aim is to give the creator the possibility to insert in his pages the indexing information to be used by the robots. Moreover these metdata should be inserted in every page leading to both a work overload for the creator and some noise in the answers. Currently, meta tags have no standardization. There is no conscensus of meaning and the users unintentionnaly or deliberately can circumvent the use of these form of metadata. Recently, research has showed a great interest on using the metadata to improve the description of electronic documents and to standardize the exchange of this metadata. For example, applications have been developped to access digital libraries through the Internet, combining text and structured fields (author, abstract, title...)

indexing within a distributed architecture (see [Witten 1996]), but have not been extended to the whole WWW [Cifford 1997]. Currently, metadata have been formalized to add semantics only on pages or data, (see references as MCF [Netscape 1997], [Luke 1997],[DC 1995]).

3. WWW documents

The WWW pages are linked in a ``flat" way, so there is neither hierarchical organization nor overall structured information. HTML structures the display of pages, but provides very little information about the content of pages or collection of pages; it is precisely this structured content we want.

3.1 What are WWW documents?

Consider the following analogy :
A librarian needs a list of words to describe the books he/she has to catalogue and a classification (for example the Decimal Dewey classification or DDC) in order to put the book in its right place on the shelves. What we call a document in this context is in general a paper document, which can be easily identified as a physical unit of information.

Let us consider now a WWW document. This kind of document is electronic and has no clear physical boundaries. For example, is a WWW page considered as a document [White 1996]? In fact, there is a great diversity in the granularity of information found on WWW servers. You can find very short and independant pages, considered as documents by themselves, and pages linked to other pages that form a cluster or another document.
We give a new definition of a *WWW document* : a document is a collection of pages or documents, created by an authority (e.g author or organization).

3.2 Describing WWW documents

We introduce now the notion of a context around a page or a document : Each page or document has a context associated with. Let's go back to the library example, and consider a paragraph of a book. This page takes place in a section, within a paragraph, within the book itself. Each level of organization of the book gives a context, from the highest level (root is the book) to the next level in the hierarchy up to the page level. These contexts may inherit from their father context, depending if the attributes are dynamic or static (this will be defined later). The contexts are composed of both pieces of information about the content of the document or external to that content. For instance, at the book level the author name, the publication date and the title are outside its semantical content.
So we suggest to make explicit these contexts and their attributes within documents, The following figure shows how the WWW pages can be organized into clusters called documents, one page can belong to different documents, therefore bounded to different contexts (here represented by the subject attribute).

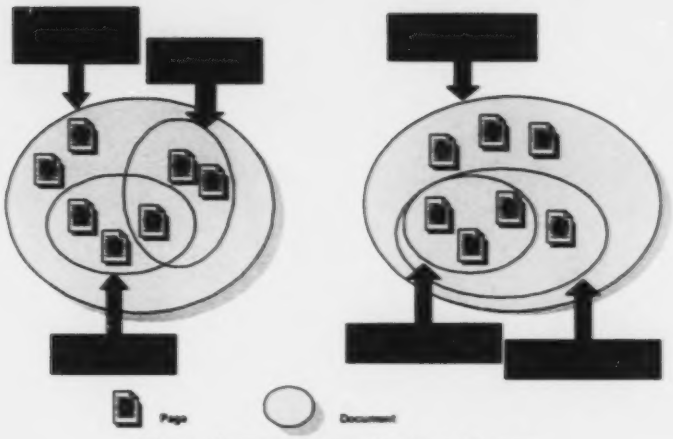

Figure 1: Several organizations of pages

Contexts are in fact represented by what we called metadata in the precedent section. Introducing metadata within a document rather than within a page has several benefits :

* describing the logical unit of information at a correct level of granularity ; enabling the author of the documents or experts to explicit the context associated with the documents;

* offering a better control over the structure of the WWW by making explicit the organization of documents;

* enabling flexibility of clusterization of pages, which correponds to several organizations of the same collection of pages;

* avoiding to duplicate the same information in all the pages which belong to the same document; therefore lightening the author work load.

The contexts have different attribute types : *static* and *dynamic*. Static means that the metadata is local to the context, then it will not be propagated along the tree of documents. Conversely, dynamic attributes means that they can be forwarded along the hierarchy. For example, coming back to the [Fig. 1], we can see that for the subject attribute, the word *Pollution* specializes *Environment*, and *Water Pollution* specializes *Pollution*.. *Environment* is forwarded along*water pollution* and *water* by inheritance.

Each author may have a certain view of the organization of the information within his pages and is responsible for the creation of the corresponding documents descriptors, or metadata. Each document contains a collection of pages or documents and is described by a set of attribute-value pairs. These metadata are based on XML for the syntax and on MetaData Dublin Core for the semantics. This is an example of the document metadata:

```
<XML>
<environment-document>
        <identifier>#env001 </identifier>
        <subject>
                <scheme>
                        <thesaurus_GEMET>    Environment    </thesaurus_GEMET>
                        <DDC> 333.7 </DDC>
                </scheme>
        </subject>
        <title> Environment server </title>
        <author>
                Bich-Liên Doan and Michel Beigbeder
        </author>
        <description> enterprise and teaching server about environment. </description>
        <subject>    Environment, pollution, water treatment </subject>
        <relation>
```

```
                <identifier-scheme> URL </identifier-scheme>
                        <type> contains </type>
                        http://www.emse.fr/ENVIRONMENT/environment.html,
                        <type> child </type>
                        http://groseille.emse.fr/ENVIRONMENT/files/d1.txt,
                        http://groseille.emse.fr/ENVIRONMENT/POLLUTION/d2.txt,
                        http://groseille.emse.fr/ENVIRONMENT/TREATMENT/d3.txt
        </relation>
</environment-document>
</XML>
```

The basic element for the representation of descriptors of documents is the inclusion relationship with other documents or pages. These relationships are represented with the relation (type = child, type = contains) attribute. With this attribute, it is possible to represent hierarchies of documents.
The subject attribute is another important one. The value of this attribute can be described with the DDC for instance. With such a scheme, any ambiguity of terms is avoided, but the author can describe freely by keywords terms as well with the subject attribute without any scheme. Let us come back to our three first requests :

```
Q1 = environment
Q2 = water treatment
Q3 = water treatment + environment
```

In contrast with the classical search tools, we produce the following results : Q1 returns the documents described by the *environment* subject. Q2 gives the set of pages containing *water treatment* embedded within *medicine* or *environment* documents.
Q3 returns the pages containing *water treatment* in the *environment* context.
We can now suggest how the use of metadata associated with documents can be helpful in a cooperative architecture of search tools.

4. Specialist and Generalist robots

4.1 Definitions

* Universal Robots : Universal robots currently exist on the WWW (eg. Alta Vista, HotBot, Lycos), although most of them include complementary topical search, they are keeping on their well-known function of indexing whole pages of the WWW.

We define two further kinds of robots : robots for general purpose tools and robots for specialist tools.

* Generalist Robots : They create an overview of the WWW by achieving two functions :

o first, they collect the metadata of whole sites on the WWW and they index descriptors of WWW documents,

o second, they manage an acquaintance database of services (addressed by other generalists and specialists) in order to route the queries towards the right services.

* specialist robots : They have knowledge in a particular domain. They use the metadata to decide if they are interested or not in exploring subtrees and indexing pages.

Our concept is based on high interactions between the entities defined above which form a specific structure. Each of the specialist robots is described by its own metadata (like a document) and can be requested by other specialists or generalists; the hierarchy of documents is then extended to the hierarchy of specialists. The robots are able to cooperate using the metadata specified in the last section. The advantage of this structure is that specialists or generalists may use metadata instead of the documents themselves for building their indices, thus reducing the network load. In this case, it is possible to improve answers to general queries (those that give thousands of answers): if a query generates 10000 URLs located on 100 sites, it is probably better to return the metadata associated with these 100 sites

rather than a (poorly) ordered list of 10000 URLs. Such non-specific queries should be addressed to the generalists. Reciprocally, well-defined queries should be addressed to specialists.
In this section, we will detail what exactly are the roles handled by the entities in our structure.

4.2 Roles

4.2.1 WWW sites providing metadata

A WWW site stores a set of WWW pages. If metadata is embedded within an existing hierarchy of documents then a site may provide :

* one or serveral organizations of pages and documents
* metadata describing pages and documents, that specify :
  - o One classification scheme (e.g DDC), or a thesaurus if used
  - o The Meta Dublin Core fields, in addition with the number of pages or volume of the site.These metadata may be stored within the site, at the root place or may be a reference to a URL stored somewhere else, for example maintained by a specialist.
* a fingerprint of others specialists which have already indexed it.
* a standard robot.txt file

4.2.2 Specialists

A specialist is created when needed to provide an identification, a description of his domain. It may contact a generalist to ``push" information about itself. Its main function consists in gathering and indexing information about its specific domain, and more precisely :

* collecting metadata according to its subject (for example environment),
* indexing HTML pages associated with the knowledge domain,
* storing and managing metadata and indexing pages (updating, deleting data...),
* routing requests to generalists if it cannot answer the query
* publishing its own metadata, to be later indexed by other robots.
* answering by giving either pages or summaries of documents, so that the user can navigate the hierarchy of contexts linked to the retrieved documents.

4.2.3 Generalists

* collecting metadata from one or several specialists while maintaining references to these specialists
* directly collecting the metadata from the WWW sites
* routing the refined queries to the adequate specialists or providing summarized responses to general user requests,
* adopting a political decision to collect all the metadata on the internet or not.

4.3 Architecture

Suppose we have two specialists S1 and S2 dealing respectively with environment and pollution. Consider now the W1 site which fills S1 and S2 in environmental data, with metadata and pages. G knows S1 and its knowledge domain, whereas S1 knows S2 which is more specific than itself.

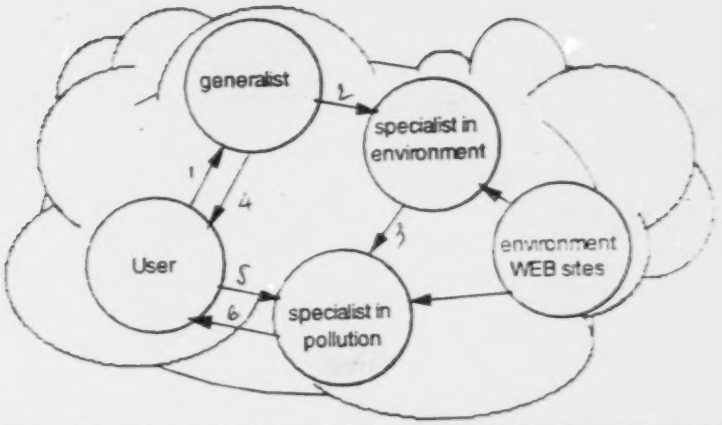

Figure 2: Interactions between specialists and generalists to answer a user need

* The user asks a generalist Q3. Q3 is translated to subject = environment. keywords = treatment + water.
* The generalist finds one specialist in the environment area. He contacts S1 and transmits the request.
* S1 knows another specialist, S2, whose area is specific to ``pollution of water". He transmits Q3 to S2.
* S1 gives the user a collection of documents and pages he retrieves from his local database and shows the context of his responses with the description of S2 included.
* S2 gives the hierarchical tree of concepts to the user.
* The user requests S2 for more detailed information.
* S2 searches his database and gives results to the user.

5. Conclusion

Defining structured metadata embedded within the documents should be used for organizing information and improving the construction of general indices. Here we have defined a scalable architecture which offers the present search tools the ability to index quickly and with better control. Our structure has the following advantages:

* Decreased consumption of the bandwidth. Robots exchange indices and may only index summaries of documents;
* More relevant answers. The contexts attached to the documents are hierarchically organized, involving interpretation and analysis of the structure and the content of the server;
* Distributed indexing. Specialist robots are focusing the information upon one particular domain;
* A self-configuring system. Specialist and generalist robots are discovering metadata from each other.

We are currently implementing a prototype of specialists and generalists for an environmental application project.

6. References

[Witten 1996] "Compression and Full-Text Indexing for Digital Libraries", Ian H.Witten, Alistair Moffat, Timothy C. Bell. DL 1994: Newark, NJ, USA. Selected Papers. Springer 1995, ISBN 3-540-59282-2

[Clifford 1997] Lynch, Clifford. Searching the Internet. Scientific American, March 1997

[Netscape 1997] "Meta Content Framework Using XML ", NOTE-MCF-XML , Netscape Communications Corporation, 06 June 1997, http://www.w3.org/TR/NOTE-MCF-XML-970624

BEST COPY AVAILABLE

[Luke 1997] "SHOE", Sean Luke, Lee Spector, David Rager, and Jim Hendler. In Proceedings of First International Conference on Autonomous Agents 1997, AA-97

[DC 1995] OCLC/NCSA Metadata Workshop: The Essential Elements of Network Object Description. March 1-3, 1995

[White 1996] "WEB document engineering", 5th International WWW conference, tutorial notes, Bebo White. May 6-10, 1996.